tussen R^2_A en R^2_B: is de toename te wijten aan de toevoeging van predictor j? Of is het toevallig?

Om f^2 te berekenen, hoeven we R^2_A en R^2_B te bepalen. Model A is misschien een bestaand model dat al empirisch onderzocht werd en je kent dus misschien R^2_A. Of je gaat misschien de proportie verklaarde variantie van een gelijkaardig model gebruiken als proxy voor R^2_A.

Voor R^2_B kan je waarschijnlijk niet rekenen op vroeger onderzoek, maar je gaat een realistische toename ($R^2_B - R^2_A$) beschouwen.

Koopgedrag 2 Je beschikt over een model met 4 predictoren om het koopgedrag van consumenten te voorspellen. Dit is model A, met 25% verklaarde variantie. Je wil nagaan of X_5 ook een predictor is van Y. Elke van de vier predictoren X_1 t.e.m. X_4 verklaren (gemiddeld gezien) $0.25/4 = 6.25\%$ van de variantie van Y. Je vermoedt dat X_5 een zwakkere predictor is, maar niet super zwak (anders wil je hem niet in model B opnemen). Je wenst dus een toets uit te voeren om 3% toename in verklaarde variantie te kunnen detecteren.

Nu kan je f^2 berekenen: $f^2 = 0.03/(1 - 0.28) \approx 0.042$. En we gebruiken nu `pwr.f2.test` met `u` $= p - k = 5 - 4 = 1$ en `power` $= 0.9$.

```
> pwr.f2.test(u=1, f2=0.042, power = 0.9)

     Multiple regression power calculation

              u = 1
              v = 250.1129
             f2 = 0.042
      sig.level = 0.05
          power = 0.9
```

We vinden `v` $= n - p - 1 = n - 5 - 1 \approx 250$. Dus $n \approx 250 + 5 + 1 = 256$. Bijgevolg, indien je een steekproef van 256 individuen trekt en indien het model B met 5 predictoren 3% van de variantie van Y kan verklaren boven wat model A al verklaart, dan zal je dit detecteren (model A verwerpen) met kans 90%.

9.9 Controle van modelassumpties: de functie `plot`

De assumpties van meervoudige lineaire regressie zijn dezelfde als die van enkelvoudige lineaire regressie: de Gauss-Markov assumpties en de normaliteitsassumptie[6].

Stel dat je een lineaire regressie met `R` hebt uitgevoerd en dat de resultaten in het object `myLM` gestopt zijn. Het commando `plot(myLM)` gaat vier diagrammen tekenen die ons helpen om de modelassumpties te checken. Telkens als je

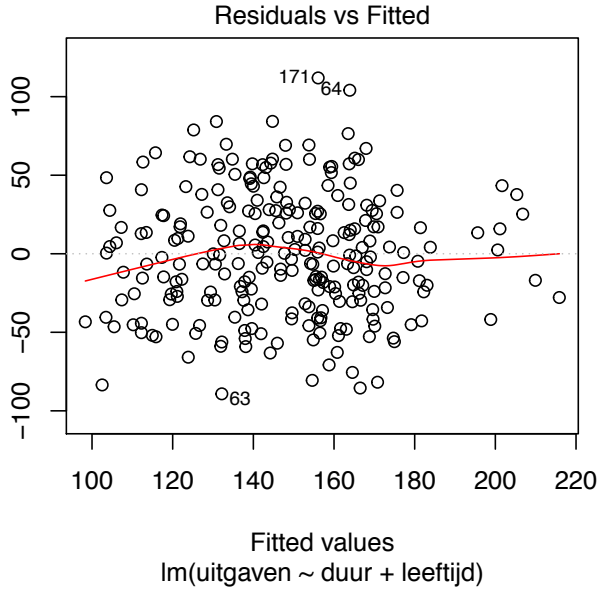
[6] Normaliteit van de residuen, niet van de predictoren of van de afhankelijke variabele

`Enter` of `Return` drukt, verschijnt een nieuw diagram. We gaan ze één per één beschrijven a.d.h.v. het gezondheidsvoorbeeld.

```
> LM <- lm(uitgaven ~ duur + leeftijd, data = gezondheid)
> plot(LM)
Hit <Return> to see next plot:
```

9.9.1 `Residuals vs fitted` — Gauss-Markov 1

Het eerste diagram is de "`Residuals vs fitted`" plot (Fig. 9.11). Op de hori-



Figuur 9.11: `Residuals vs fitted` plot — gezondheidsvoorbeeld

zontale as vind je de predicties (Engels: fitted) en op de verticale as, de residuen. Dit diagram wordt gebruikt i.p.v. het klassieke spreidingsdiagram[7] om de residuen te analyseren. Dit diagram wordt niet in detail uitgelegd; we zien wel hoe het gebruikt wordt.

Op Fig. 9.11 vind je een rode curve. Elke punt op deze curve representeert de schatting van de voorwaardelijke verwachting van ε_i. Elke punt op deze rode curve is het gemiddelde van de corresponderende verticale snede. De eerste Gauss-Markov assumptie stelt dat $E(\varepsilon_i) = 0$. Dit impliceert dat de voorwaardelijke verwachting van de residuen nul is. De rode curve moet dus min of meer horizontaal zijn, op hoogte 0 (aangeduid door de horizontale stippellijn). Op Fig. 9.11 zien we geen duidelijke afwijking t.o.v. de stippellijn. We mogen dus de eerste Gauss-Markov assumptie aanvaarden.

[7] Het klassieke spreidingsdiagram kan niet getekend worden bij meervoudige lineaire regressie indien het aantal predictoren groter dan 2 is.

Op Fig. 9.11 vind je ook een paar punten met een getal ernaast. Dit zijn punten die door `R` als outliers of speciale punten geïdentificeerd worden. Het getal naast het punt is het nummer van de corresponderende rij in het data frame. Het is aangeraden om die punten afzonderlijk te bekijken.

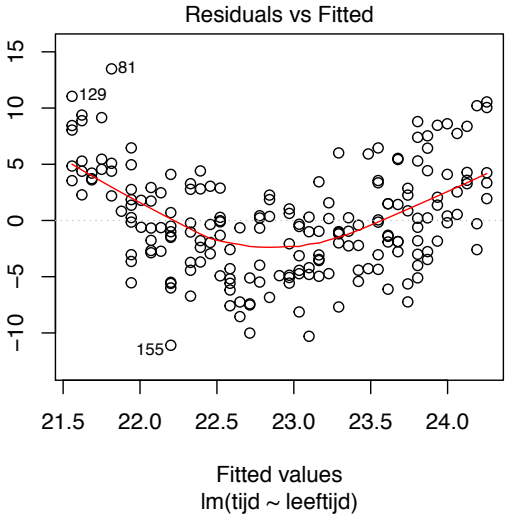
```
> gezondheid[c(63,64,171),]
    geslacht duur uitgaven leeftijd
63         M   22       43       30
64         V   33      268       40
171        V   31      268       36
```

Je kan verifiëren of er geen tikfout is of een ander probleem.

Schending van de 1ste Gauss-Markov assumptie Op Fig. 2.7 hebben we gezien dat het verband tussen `tijd` en `leeftijd` in het data frame `sportData` curvilineair is. Laten we dit analyseren met de functie `lm`.

```
> LM.tijd.leeftijd <- lm(tijd ~ leeftijd, data = sportData)
> plot(LM.tijd.leeftijd)
Hit <Return> to see next plot:
```

De eerste plot wordt in Fig. 9.12 weergegeven. We zien dat de schattingen van



Figuur 9.12: `Residuals vs fitted` plot — sportData

de voorwaardelijke verwachtingen helemaal niet constant zijn: de rode curve heeft min of meer de vorm van een parabool. De eerste Gauss-Markov assumptie wordt dus waarschijnlijk niet voldaan. Dit wijst aan dat het verband tussen de twee variabelen niet lineair is (zie ook Fig. 8.9).

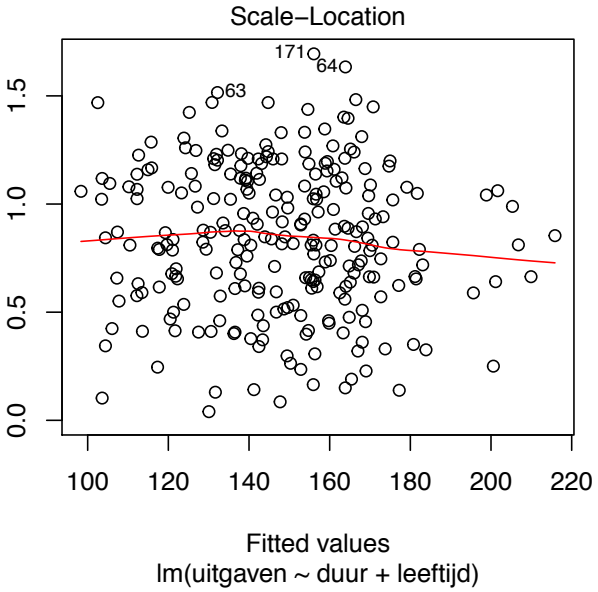
103. Teken de `Residuals vs fitted` *plot voor het lineair model dat het gewicht van de baby verklaart m.b.v. van* `gestation`, `parity`, `age`, `wt.1`, `number`, `dage` *en* `dwt`. *Is de eerste Gauss-Markov assumptie in orde?*

9.9.2 `Normal Q-Q` — Normaliteit

Het tweede diagram dat door het commando `plot(LM)` wordt getekend, is de normale qq-plot. We hebben dit diagram al vaak besproken. We maken toch een opmerking. Als je dit diagram met het commando `plot(LM)` tekent (en niet met `qqnorm(residuals(LM))`), dan worden opnieuw een paar punten door `R` geïdentificeerd als outliers. Het is aangeraden om die punten afzonderlijk te bekijken.

9.9.3 `Scale-Location` — Homoscedasticiteit

Het derde diagram dat door het commando `plot(LM)` wordt getekend, is de "`Scale-Location`" plot (Fig. 9.13). Dit diagram wordt niet in detail uitgelegd; we zien wel hoe het gebruikt wordt. Op Fig. 9.13 vind je een rode curve.

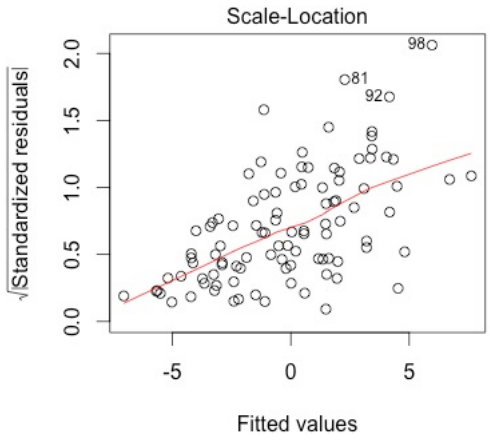


Figuur 9.13: `Scale-Location` plot — gezondheidsvoorbeeld

Elke punt op deze curve representeert de schatting van de vierkantswortel uit de voorwaardelijke variantie van Y. De tweede Gauss-Markov assumptie stelt dat de voorwaardelijke variantie van de residuen constant is. De rode curve moet dus min of meer horizontaal zijn. Op Fig. 9.13 zien we dat de rode curve min of meer horizontaal is. We mogen dus de tweede Gauss-Markov assumptie aanvaarden.

Op Fig. 9.13 vind je ook een paar punten met een getal ernaast. Dit zijn punten die door `R` als outliers geïdentificeerd worden. Het is aangeraden om die punten afzonderlijk te bekijken.

Schending van de homoscedasticiteitsassumptie Fig. 9.14 geeft de derde plot ("`Scale-Location`" plot) weer van een fictief voorbeeld. We zien dat de



Figuur 9.14: `Scale-Location` plot — fictief voorbeeld

schattingen van de voorwaardelijke variantie helemaal niet constant zijn: de rode curve stijgt. De tweede Gauss-Markov assumptie wordt dus waarschijnlijk niet voldaan.

104. Teken de `Scale-Location` *plot voor het lineair model dat het gewicht van de baby verklaart m.b.v. van* `gestation`, `parity`, `age`, `wt.1`, `number`, `dage` *en* `dwt`. *Is de tweede Gauss-Markov assumptie in orde?*

9.9.4 `Residuals vs Leverage` — Invloedrijke punten

Het vierde diagram dat door het commando `plot(LM)` wordt getekend, is de "`Residuals vs Leverage`" plot (Fig. 9.13). Dit diagram wordt niet gezien.